Application of Information Technology ■

A Software Tool for Removing Patient Identifying Information from Clinical Documents

F. JEFF FRIEDLIN, DO, CLEMENT J. MCDONALD, MD

Abstract We created a software tool that accurately removes all patient identifying information from various kinds of clinical data documents, including laboratory and narrative reports. We created the Medical De-identification System (MeDS), a software tool that de-identifies clinical documents, and performed 2 evaluations. Our first evaluation used 2,400 Health Level Seven (HL7) messages from 10 different HL7 message producers. After modifying the software based on the results of this first evaluation, we performed a second evaluation using 7,190 pathology report HL7 messages. We compared the results of MeDS de-identification process to a gold standard of human review to find identifying strings. For both evaluations, we calculated the number of successful scrubs, missed identifiers, and over-scrubs committed by MeDS and evaluated the readability and interpretability of the scrubbed messages. We categorized all missed identifiers into 3 groups: (1) complete HIPAA-specified identifiers, (2) HIPAA-specified identifier fragments, (3) non-HIPAA–specified identifiers (such as provider names and addresses). In the results of the first-pass evaluation, MeDS scrubbed 11,273 (99.06%) of the 11,380 HIPAA-specified identifiers and 38,095 (98.26%) of the 38,768 non-HIPAA–specified identifiers. In our second evaluation (status postmodification to the software), MeDS scrubbed 79,993 (99.47%) of the 80,418 HIPAA-specified identifiers and 12,689 (96.93%) of the 13,091 non-HIPAA–specified identifiers. Approximately 95% of scrubbed messages were both readable and interpretable. We conclude that MeDS successfully de-identified a wide range of medical documents from numerous sources and creates scrubbed reports that retain their interpretability, thereby maintaining their usefulness for research.

■ **J Am Med Inform Assoc.** 2008;15:601–610. DOI 10.1197/jamia.M2702.

Introduction

Since the practice of medicine began over 25 centuries ago, physicians have had a duty to protect a patient's privacy. The Hippocratic Oath states: "I will respect the privacy of my patients, for their problems are not disclosed to me that the world may know."[1] Relatively recent technological innovations such as the World Wide Web, electronic medical record (EMR) systems, and increased connectivity between disparate medical institutions, although improving medicine by facilitating the sharing of patient data, have increased the challenge of protecting patient privacy.[2,3] In 1996, Congress passed the Health Insurance Portability and Accountability Act (HIPAA),[4] calling for standards to protect individuals' health information. According to HIPAA regulations, protected health information (PHI) is individually identifiable health information transmitted by electronic media, maintained in electronic media, or transmitted or maintained in any other form or medium. The Department of Health and Human Services in turn issued the Privacy Rule, which established national standards to protect such information.[5]

Privacy Rule regulations apply to protected health information. According to the HIPAA Privacy Rule, when patient records are de-identified by removing all HIPAA-specified patient identifiers (Figure 1), the de-identified data set is no longer considered protected health information. The HIPAA states that before de-identified data are released, they must be verified as de-identified using either statistical methods or by manual review. The HIPAA allows entities to release data that have been de-identified without obtaining an authorization and without further restrictions on use or disclosure. As EMRs become more common and widespread and patient data become increasingly accessible to researchers, the need for automated de-identification of medical data will become more acute.

Clinical data can be de-identified by removing all of the 19 HIPAA specified identifiers from a clinical document (Figure 1). Several systems have recently been described that remove patient identifiers from pathology reports[6–9] and from databases.[10–12] Results of these systems varied between 0.82 to 0.98 sensitivity and 0.75 to 0.99 specificity. These systems used various methods such as natural language processing,[12] complex rule sets,[10] and specialized dictionaries or name lists[6,9] to perform de-identification. Four systems were specially designed for and tested on pathology reports exclusively.[6–9] Only one study[7]

Affiliations of the authors: Regenstrief Institute, Indianapolis, IN. Dr. McDonald is currently at the Lister Hill Center, Bethesda, MD.

The authors thank the Regenstrief research personnel who assisted with this project.

All patient names cited in this paper, although analogous in format to the original names, are fictitious.

Correspondence: Jeff Friedlin, DO, Regenstrief Institute, Inc., Medical Informatics, Health Information and Translational Sciences (HITS) Building, 410 West 10th Street, Suite 2000, Indianapolis, IN 46202; e-mail: <jfriedlin@regenstrief.org>.

Received for review: 12/20/07; accepted for publication: 05/30/08.

1. Name
2. All geographic subdivisions smaller than a state (street address, city, county, precinct)
Note: zip code or equivalents must be removed, but can retain first 3 digits if the geographic unit to which the zip code applies if the zip code area contains more than 20,000 people
3. For dates directly related to the individual, all elements of dates, except year. (date of birth, admission date, discharge date, date of death)
4. All ages over 89 or dates indicating such an age
5. Telephone number
6. Fax number
7. Email address
8. Social Security Number
9. Medical Record Number
10. Health Plan Number
11. Account Numbers
12. Certificate or license numbers
13. Vehicle identification/serial numbers including license place numbers
14. Device identification/serial numbers
15. Universal Resource Locators (URL's)
16. Internet Protocol addresses (IP's)
17. Biometric Identifiers
18. Full face photographs and comparable images
19. Any other unique identifying number, characteristic or code

Figure 1. Nineteen patient identifiers that require removal for de-identification per HIPAA regulations.

mentioned that its system was tested in de-identifying other types of clinical documents, but did not report the system's accuracy.

Both De-Id[13] and Deidentify[14] are commercial automated de-identification systems that process medical texts. The De-Id is a standalone tool that uses rule sets, heuristics, and supplemental dictionaries to remove PHI from medical documents. It replaces patient identifiers with specific tags in the form of offsets and proxies in an attempt to preserve the usability of data. Deidentify is a software component that can be used to create Java and .NET programs that remove PHI from medical documents. It can also be configured to remove specific elements of protected health information as well as extended using regular expressions.

In 2007, the American Medical Informatics Association (AMIA) sponsored an automated de-identification challenge as part of the i2b2 (informatics for Integrating Biology to the Bedside) project.[15] Documents used in the challenge were discharge summaries with both a training set (669 reports) and test set (220 reports) distributed. The PHI in these reports were manually replaced with realistic surrogates. Seven teams of developers participated and used de-identification systems ranging from pure rule-based systems and pure statistical learning systems to hybrid systems using both methodologies. Results showed that statistical learning systems using rule templates as features performed best, followed by hybrid systems of rules and machine learning. The best-performing system applied 2 named-entity recognizers complemented with regular expressions to the de-identification task.[16] The second-best-performing system also applied an interative named-entity recognizer augmented by decision trees with local features and dictionaries.[17] The set of documents used in this challenge was small compared with other de-identification studies.

We developed the Medical De-identification System (MeDS), a computer software tool for de-identifying Health Level Seven (HL7) messages and narrative text documents. Here we describe MeDS and report on the results of 2 separate evaluations performed to measure its success de-identifying a wide variety of reports from multiple sources.

Materials and Methods

The HL7 version 2.5 message standard[18,19] is the most common method of transmitting patient information, including administrative data, laboratory reports, and free text clinical reports. It is a highly structured messaging system that clearly defines what information is required in each message and where that information is placed. We used HL7 observation messages (specifically unsolicited transmission of an observation [ORU][19] messages) as the primary input into our scrubber system for this study. After obtaining approval from the Institutional Review Board of the Indiana University Medical center, we randomly selected samples of HL7 messages transmitted to the Regenstrief Medical Record System (RMRS),[20] which is the database for the Indiana Network for Patient Care (INPC).[21] The INPC is a local health information infrastructure that includes information from 5 major hospital systems (15 separate hospitals) and more than 100 clinics and day surgery facilities distributed throughout Indianapolis and the surrounding counties. Together they generate 165,878 inpatient admissions, 450,000 emergency room (ER) visits, and 2.7 million outpatient visits per year.[21]

For our initial evaluation of our software, we randomly collected 2,400 HL7 messages, 200 consecutive messages from each of the 10 sources in Table 1. All were produced in 2005. Narrative sources A and B listed in Table 1 are HL7 messages containing various narrative reports. Table 2 shows the distribution of the types of reports in the 2 sources. To more intensely analyze admission notes because they are highly variable and the most open-ended, we separately extracted an additional 200 consecutive messages containing only admission notes from 1 HL7 message producer. We also randomly collected a consecutive series of 200 HL7 messages sent to the RMRS regardless of source.

Based on the results of our software's performance in scrubbing this initial test set, we modified the software to improve accuracy. After performing these modifications, we conducted a second evaluation in which we collected an additional 7,193 HL7 pathology messages randomly chosen from all messages produced during the month of September 2006. We used pathology messages exclusively for this second evaluation because they contained the richest variety and frequency of HIPAA-specified identifiers.

All messages collected were of the ORU type, meaning that each message conveyed some type of patient result. All of the laboratory message producers listed in Table 1 are

Table 1 ■ Sources of Collected HL7 Messages for the First Evaluation

Laboratory A	Reference laboratory located outside the Indiana Network for Patient Care
Laboratory B	Laboratory supplying inpatient and outpatient results to hospital system #1
Laboratory C	Laboratory supplying inpatient and outpatient results to hospital system #2
Laboratory D	Laboratory supplying outpatient results to a heart hospital in hospital system #2
Laboratory E	Laboratory supplying inpatient and outpatient results to hospital system #3
Laboratory F	Quest laboratory supplying inpatient and outpatient results to hospital system #2
Laboratory G	Laboratory supplying inpatient and outpatient results to hospital system #4
Pathology	CoPath pathology transcription service supplying reports to hospital system #5
Narrative A	Transcription service supplying inpatient reports to hospital system #4
Narrative B	Transcription service supplying inpatient and outpatient reports to hospital system #5

INPC-associated inpatient and outpatient laboratories, except for Producer A, which is a reference laboratory located outside of the INPC. The narrative report samples included a variety of inpatient and outpatient narrative reports, such as admission notes, discharge summaries, consultations, operative reports, clinic notes, and endoscopy reports.

De-identification Software

MeDS takes clinical documents in a plain text file as input and produces a text file containing the same clinical document in a de-identified (or scrubbed) form. The MeDS is especially tuned to process text messages in the HL7 format, but can easily be modified to accept other institution-specific formats. For this study, we used HL7 messages as the input. All HIPAA-specified identifiers listed in Figure 1 are removed from the message, with the exception of biometric identifiers and full-face photographs, which are not contained in these text-only messages. In addition to HIPAA-specified identifiers, MeDS also scrubs health care provider names (including nurses, physicians, and laboratory personnel), other identifiers of health care personnel (office telephone and fax numbers, office addresses, etc.), and health care institution names (including medical office corporate names, hospital or hospital system names, etc.). We scrubbed such identifiers to safeguard the privacy of the health care parties involved. MeDS is also stricter than what HIPAA requires by scrubbing all references to ages and times. MeDS is written in the JAVA programming language and scrubs reports through a series of 4 scrubbing processes diagramed in Figure 2. We initially developed MeDS by testing its scrubbing performance on a small test set of 750 HL7 messages, completely independent of our study set.

Table 2 ■ Distribution of Report Types Contained in Narrative Sources A and B

Report Type	Number
Narrative source A	
History and physical	56
Discharge summary	54
Consult	26
Operative report	26
Endoscopy report	16
Cardiac catheterization report	8
Other	14
Narrative source B	
Return clinic visit note	82
Operative report	54
Discharge summary	31
New patient evaluation	25
Other	8

MeDS takes advantage of knowledge of well-labeled patient identifiers present in report headers and in lead segments of an HL7 message such as the Patient Identification (PID) segment, Patient Visit (PV1) segment, Next of Kin (NK1) segment, etc. It uses the information from these segments to find the same identifiers in the unlabeled body of narrative reports or comment sections of a laboratory report. MeDS saves this labeled information as discrete data elements (i.e., a patient's first, middle, and last names are saved separately) for later use in the scrubbing of the payload sections (defined below). The PID segment, for example, may carry the patient's name, address, social security number, alias names, telephone number, etc. The NK1 segment can contain family member names. In HL7 ORU messages, the payload is in 2 kinds of segments: the Observation (OBX) segment and the Notes and Comments (NTE) segment.[18] The OBX segment reports either discrete observations or narrative report sections, and the NTE segment typically contains comments related to the OBX segments. Both segments represent the payload, the clinically useful information in the message.

To remove numeric identifiers (such as medical record numbers, social security numbers, telephone numbers) MeDS uses a series of 11 regular expressions (available on request from J.F.). For example, one regular expression removes any string (any group of characters with no white space) with over 4 consecutive digits. This may cause some over-scrubbing of laboratory values 10,000 or greater, but we are willing to accept the loss of some clinical information to ensure accurate scrubbing of identifiers. Another regular expression detects the 3-digit, 2-digit, 4-digit pattern of a social security number. MeDS similarly removes dates using a series of 10 regular expressions. We discovered we needed this number of regular expressions to account for all variations in which a date could be written (Figure 3). MeDS uses 10 other regular expressions to remove all place names, address patterns, references to location, etc. For example, one regular expression searches for a digit-text-street pattern (i.e., 127 Main Street) to remove likely addresses. Another regular expression detects state names or abbreviations, and yet another detects other location-digit patterns such as Suite 222, Building 4, or Room 137. We use 2 regular expressions to detect time patterns (7:13 AM), and one is used to detect references to age (92-year-old). An additional 7 regular expressions are used to detect other identifier patterns such as e-mail addresses (text@text), provider names ("send copy to *name*", "report dictated by *name*") etc.

The MeDS uses different methods to detect and scrub proper names in a report. It uses word lists and algorithms similar

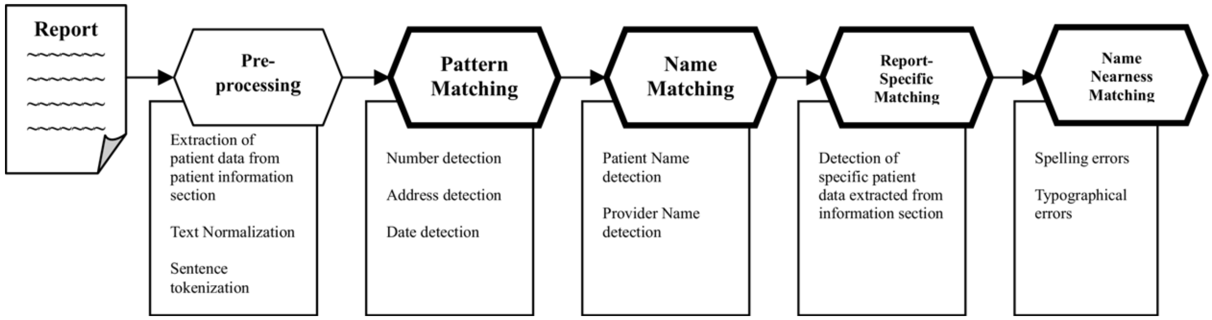

Figure 2. Processing schema of the de-identification software.

to that described by Thomas et al.[6] The MeDS uses 2-word lists during this process: (1) a list of proper names, and (2) a list of clinical and common usage words (CCUW). We compiled the proper names list from 3 sources: (1) all proper names from an open-source spell-checking dictionary (Ispell),[22] (2) all health care provider first and last names in the RMRS, and (3) all names from the 2005 Social Security Death Index. We included first and last names as separate proper names, and all duplicate names were removed. We compiled the CCUW word list from 2 sources: (1) the Unified Medical Language System (UMLS) word index of words with a Source Restriction Level of 0 or 1, and (2) the word list from the Ispell dictionary (excluding proper names, which are capitalized). We performed additional modifications of the lists described by Thomas et al.[6] The proper name list and the CCUW list were compared, and all words common to both lists were extracted and placed into a common word list totaling 5,801 words. One of the investigators (J.F.) manually reviewed this common word list. We compared this common word list with the Medical Subject Headings (MESH) vocabulary. If a word in the common word list was not present in MESH, we deemed the word less likely to appear in a medical document in a medical context, and it was returned to the proper name list. For example, the word angel appeared in both the proper name list and the CCUW list. Because angel is not a MESH term, we determined that this word was not likely to appear in a medical document except in the context of a proper name, so it was returned to the proper name list. We returned 3,539 such words to the proper name list. Conversely, if a word from this common word list was found in MESH, it was returned to the CCUW list. We returned 2,262 such words to the CCUW list. After editing, the proper name list contained 284,323 unique names and the CCUW list contained 303,057 words. We additionally supplemented the proper name list with the top 16,000 (ranked by population) city names from the 2000 U.S. census data. In addition to using these lists, MeDS also searches for predictive markers that likely represent proper names. These include proper name prefixes such as Mr., Mrs., and Dr., and proper name suffixes such as MD, Jr., and PhD. For example, when a common word such as white is found, MeDS examines the words surrounding it. If "Mr." precedes it, or "MD" follows it, the word 'white' is scrubbed. The algorithm for how our software uses the name lists and predictive markers is derived from Thomas et al.[6] and is shown in Figure 4.

January 23, 2006
January 23rd, 2006
Jan 23, 2006
01/23/2006
1/23/2006
01/23/06
01-23-2006
01.23.2006
20060123
20062301
200601232216 (date/time)

Figure 3. Examples of alternative date display formats found in sample messages.

We added a supplementary procedure to the scrubbing process of proper names described by Thomas et al.[6] We included a part-of-speech processor to assist with disambiguation of proper names. We used a modified version of the part-of-speech tagger in MetaMap,[23] a system of software tools for text processing produced by the National Library of Medicine (NLM). For example, the proper name list contains names that, although not commonly used words, could still appear in a medical document in a non–proper name format. Ideally, MeDS would accurately determine when a term such as *colon* is being used as a proper name and when it is being used in a medical context. We observed that when a word is a proper name in a report, it is usually not preceded by an adjective, determiner, or pronoun. For example, for the phrases "I examined the *pat*," "she described a *green* discharge," and "she *may* have cancer," all of the italicized words are in the proper name list but are not proper names in their respective contexts. In all of the phrases, the words in question are preceded by pronouns or determiners, which make them unlikely to be proper names in this sentence. To minimize the risk of missing true proper names, the part-of-speech processor is called only when no other evidence of a proper name exists; i.e., no surrounding known proper names and no prefixes/suffixes likely to be associated with proper names are present. The MeDS uses an additional process to detect and scrub specific message-related identifying data, such as patient names and addresses. Using

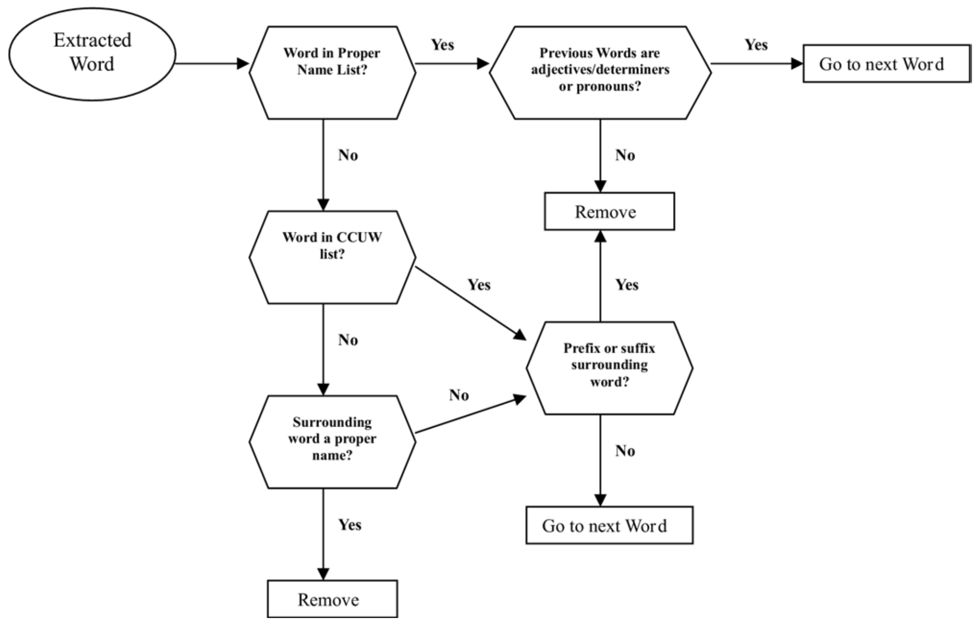

Figure 4. Algorithm used by the name scrubbing process.

identifiers extracted earlier from the lead section of the HL7 message, MeDS scrubs matching patient identifiers found in the payload section of the message. We use the direct knowledge of names and addresses that are known to be the patients' (or that of some closely related third party) in this scrubbing process. The MeDS also uses this information to detect typographical errors and name variants in a message. Misspelling of words and names can cause havoc with de-identification systems[7–9] and natural language processing systems.[24,25] Therefore, we added another process that uses a text string nearness algorithm. The algorithm is a modified version of the one reported by Friedman et al.[26] Every word in the payload section of a clinical document is processed by this algorithm for nearness to the patient's known first, middle, and last names. The algorithm compares the target word with the patient's name using a combination of the longest common string and the common ordering of letters, and outputs a nearness score between 0 and 1. The higher the score, the greater the similarity between the 2 words, with a score of 1.0 being a perfect match. Based on the threshold we use currently in a related patient linker system, we consider a score of 0.70 or greater as a match and warrants scrubbing. Thus, the software can accurately detect the name "Smith" when it is misspelled "Smithe," "Smit," or "Ssmith." It can also detect "Johnny" when the patient's known name is "John." If a patient's name is similar to a common word, such as "Brandon" and "random." then over-scrubbing errors likely will occur. We accept a degree of over-scrubbing to minimize the risk of under-scrubbing a misspelled patient name.

Evaluation

We evaluated the performance of MeDS by comparing the results to human review (by one of the physician authors) of the scrubbing performance. He built this gold standard in 2 phases. He hand-reviewed a mixed set of 200 messages (50% laboratory messages, 50% narrative reports) before they were scrubbed. For the remaining sample of 2,200 messages, he reviewed them after they were scrubbed. We did this to assess whether the results of the evaluation process differed depending on when the review occurred (before scrubbing vs. after scrubbing). For the second study of 7,193 pathology messages, the review of the accuracy was done by the same physician after the software scrubbing. In all messages, the reviewer recorded whether the identifier was patient related (HIPAA specified) or provider related (non-HIPAA specified). To facilitate this review, the software showed the text from the original message that it removed (in brackets) to de-identify the HL7 message.

We counted MeDS's failure to remove full patient identifiers, including any of the 19 identifiers listed in Figure 1, fragments of patient identifiers (e.g., middle name initials, a partial address), and identifiers for providers, which are not forbidden by HIPAA. We also looked for other nonpatient identifiers, including health care personnel names, institution names, provider phone numbers, etc.

For each message, we calculated 3 totals: (1) the number of identifiers that MeDS identified for removal, (2) the number of instances that MeDS failed to remove identifiers (under-scrubbing), and (3) the number of instances when a word (or words) were removed that were not identifiers (over-scrubbing).

Results

Results of the First Phase

For our first phase, we collected a sample of 2,400 total HL7 messages, including 1,400 laboratory messages; 800 narrative radiology, pathology, and hospital dictation messages; and 200 mixed-source messages. The sample contained approximately 446,000 words in the payload sections (OBXs and NTEs) of the message. Because of the structure of some HL7 messages, the same identifiers may be repeated many times within 1 message. For example, a laboratory message with 12 OBX segments could contain the same laboratory-specific specimen number, date, time, etc., in each segment. This tends to inflate the measure of scrubbing success for those kinds of identifiers that occur repeatedly. To reduce this effect, we counted only the first unique identifier per

Table 3 ■ Number of HIPAA-specified and Non–HIPAA-specified Identifiers Present Per Message Source

Message Source (200 Reports Each)	Total Words	HIPAA-Specified Identifiers	Non–HIPAA-Specified Identifiers	HIPAA Identifiers Per Word	Total Identifiers
Laboratory A	35,957	0	3,020	0.000	3,020
Laboratory B	9,814	323	1,609	0.033	1,932
Laboratory C	15,356	236	3,218	0.015	3,454
Laboratory D	14,639	196	2,591	0.013	2,787
Laboratory E	32,052	233	7,214	0.007	7,447
Laboratory F	37,978	227	7,718	0.006	7,945
Laboratory G	11,230	188	1,093	0.017	1,281
Pathology	56,076	2,211	341	0.039	2,552
Transcription A	69,108	1,935	1,396	0.028	3,331
Transcription B	70,840	3,140	3,892	0.044	7,032
Admission notes	73,421	1,513	2,177	0.021	3,690
Mixed source	19,458	1,178	4,499	0.061	5,677
Total	445,929	11,380	38,768	0.026	50,148

HIPAA = Health Insurance Portability and Accountability Act.

message in our numerators and denominators. Based on physician review, we observed 50,148 total unique identifiers in the first sample of 2,400 HL7 messages (average of 21 per report, range 0 to 35), and 97% of message payloads contained at least 1 identifier.

We calculated the number of HIPAA-specified identifiers and non–HIPAA-specified identifiers for each message source (Table 3). The 2,400 HL7 messages contained 11,380 unique HIPAA-specified identifiers and 38,768 unique non–HIPAA-specified identifiers. The MeDS scrubbed 11,273 (99.06%) of the HIPAA-specified identifiers, and scrubbed 38,095 (98.26%) of the non–HIPAA-specified (Table 4). The MeDS missed no complete HIPAA identifiers; however, it did miss fragments of these identifiers. For example, in several reports MeDS missed fragments such as the street number of a patient's address, or parts of a city name (i.e., it missed the 'New' in 'New Town' but scrubbed the rest of the address). Also, MeDS occasionally missed the middle initial but scrubbed the first and last names of the patient. Importantly, we observed no instances in which MeDS missed a patient's first or last name. In several reports, fragments of a date (i.e., "03/20/05" was converted to "/20/05") were missed. We calculated the fragment proportions missed per report by dividing the total number of fragments by the number of reports. The MeDS missed an average of 0.05 fragments per message. There were 0.12 fragments missed per report in the 600 messages containing narrative reports, and 0.17 fragments missed per report in the 200 pathology report messages.

Table 4 ■ Under-scrubbing Errors Committed by Software for Each Message Type in First Evaluation

Under-scrubbing Errors	Total	Missed
HIPAA-specified Identifiers		
Laboratory (n = 1,400)	1,403	0 (0.0%)
Narrative reports (n = 600)	6,588	73 (1.1%)
Pathology reports (n = 200)	2,211	34 (1.5%)
Mixed message (n = 200)	1,178	0 (0.0%)
Non–HIPAA-specified Identifiers		
Laboratory (n = 1,400)	26,463	80 (0.3%)
Narrative reports (n = 600)	7,465	531 (7.1%)
Pathology reports (n = 200)	341	7 (2.1%)
Mixed message (n = 200)	4,499	55 (1.2%)

HIPAA = Health Insurance Portability and Accountability Act.

Regarding non-HIPAA identifiers, there were 18.90 unique identifiers per laboratory report, 12.44 per narrative report, and 1.71 per pathology report. Including both full identifiers and identifier fragments, MeDS missed 0.06 (.32%) identifiers per laboratory report, 0.89 (7.15%) identifiers per narrative report, and 0.04 (2.34%) per pathology report. All missed non-HIPAA identifiers were provider names or addresses, the majority caused by report formatting irregularities and spelling or typographical errors (we did not apply our name nearness algorithm to provider names or addresses).

The number and frequency of unique HIPAA-specified identifiers present in the reports varied depending on message type. Laboratory messages contained only 2 kinds of HIPAA-specified identifiers: dates (98%, 18.52 per report) and specimen numbers (2%, 0.38 per report). Narrative and pathology report messages contained 5 kinds of HIPAA-specified identifiers: dates (66%, 5.49 per report), patient names (31%, 2.58 per report), patient-related numbers (medical record, social security, specimen) (2%, 0.18 per report), patient addresses (0.4%, 0.03 per report), and patient ages (0.2%, 0.01 per report).

The timing of reviewer marking of the messages did not affect the results. The set of 200 HL7 mixed-source messages (marked before scrubbing) included laboratory messages (50%), narrative reports (35%), and other (15%). There were 1,178 (5.89 per report) HIPAA-specified identifiers and 4,499 (22.49 per report) non-HIPAA identifiers in this mixed sample. Our software's scrubbing performance was essentially equal to that in the sample marked after scrubbing. The software missed none of the 1,178 HIPAA-specified identifiers present in this mixed sample and missed 55 (1.22%, 0.28 per report) of the 4,499 non-HIPAA identifiers. The timing of reviewer marking of the messages did not affect the results.

To determine to what extent MeDS performance relies on data obtained from the header section of a message (mainly patient and provider names), we separately calculated the number of scrubbings performed by the patient-specific scrubber. This process performed a total of 90 scrubbings

Table 5 ■ Details of Pathology Messages Used in the Second Evaluation

Total Reports	Total Words	HIPAA-Specified Identifiers	Non–HIPAA-Specified Identifiers	Total Identifiers
7,193	1,979,102	80,418	13,091	93,509

HIPAA = Health Insurance Portability and Accountability Act.

(0.04 per message) in our set of 2,400 messages. All of these occurred in the narrative reports (pathology 7, transcription A 38, transcription B 13, admission notes 32); none occurred in the 1,400 laboratory messages. We found that 5,123 identifiers were either patient or provider names in these sources. Of these identifiers, <2% were scrubbed by this process.

We analyzed all 296 instances in which our software's name nearness scrubber replaced a word (indicating possible misspelling of a patient's name). Of 296 replacements, 15 (5%) were true patient name misspellings that would have been retained if not for this scrubber. Of these 15 misspellings, 8 were typographical errors such as no space between the previous word and the patient name, and 7 involved missing letters or letter transpositions. In 281 (95%) of the 296 nearness replacements, the match was not a patient name and was therefore an over-scrubbing error.

We found that 4,012 (8.00% of total) scrubbings were not true patient identifiers (over-scrubbed).The number of over-scrubbing errors ranged from 0 to 5 per message, with an average of 1.7 over-scrubs per message. Over-scrubbing was caused by medical terms that were mistakenly retained in the proper name list (i.e., breast) and when a laboratory test value contained more than 4 consecutive digits (i.e., platelet count of 68,000).

Results of Post-modification Phase

For our evaluation of the software status after modification, we randomly collected a test set of 7,193 pathology messages produced by 1 HL7 message producer during the month of September 2006. We took this large sample from pathology reports because pathologists often dictate names and hospital numbers of patients contained on the specimen labels and so present more patient identifiers to find. We counted the numerators and denominators the same as we did for the first sample. The payload sections in this sample of 7,193 messages contained approximately 1,900,000 words. As in the initial evaluation, we counted only the first unique identifier in each message, and separated HIPAA-specified patient and nonpatient identifiers. We observed 93,509 unique identifiers, an average of 17 per report (range 5 to 29), of which 80,418 were HIPAA-specified patient identifiers and 13,091 were nonpatient identifiers (Table 5). MeDS scrubbed 79,993 (99.47%) of the patient identifiers, and scrubbed 12,689 (96.93%) of the nonpatient identifiers (Table 6).

As in our initial evaluation, no full patient identifiers were missed. MeDS did miss 425 patient identifier fragments (2.13 per report). Of the 425 missed patient identifier fragments, 264 (62.12%, 1.32 per report) were single middle initials of patient's names, 121 (28.47%, 0.61 per report) were portions of date (day number or month number), and 40 (9.41%, 0.20 per report) were street numbers in addresses. None of these retained fragments are truly patient identifiers.

Discussion

MeDS effectively de-identified a wide variety of HL7 messages from multiple sources and is comparable to our gold standard. In our first evaluation, our software scrubbed 49,368 (98.45%) of all 50,148 unique identifiers (patient and nonpatient). It scrubbed 99.06% of all unique HIPAA-specified patient identifiers in this sample. In our second test on pathology reports exclusively, MeDS scrubbed 92,682 (99.12%) of all 93,509 unique identifiers, and 79,993 (99.47%) of the 80,418 HIPAA-specified patient identifiers. These results are better than the performance of most other reported systems. Thomas et al.[6] reported removing 7,151 (92.75%) of 7,710 names, with a system designed to scrub only proper names from pathology reports. Beckwith et al.[9] reported removing 3,439 (98.23%) of 3,499 unique identifiers from pathology reports. Gupta et al.[7] reported their software "performed extremely well" but did not quantify the results. Regarding over-scrubbing, in our first evaluation our software committed 4,012 over-scrubs, roughly 7% of the total number of found identifiers (54,160), a much smaller percent than reported by others. Beckwith et al.[9] reported 4,671 over-scrubs, roughly equal to the total number of found identifiers (4,515). Neither Thomas et al.[6] nor Gupta et al.[7] reported their over-scrubbing rates. Our study is unique in that we evaluated our software's performance in scrubbing multiple types of documents, not just a single type of report. In the recent i2b2 de-identification challenge, Uzuner et al.[15] evaluated 7 different systems and report sensitivities ranging from 0.80 to 0.96 and specificities ranging from 0.83 to 0.97. The performance of our system is similar to the best systems in this challenge, although several differences in study design prevent direct comparison. In our study, we processed nearly 9,000 reports of varying types. The i2b2 test

Table 6 ■ Under-scrubbing Errors Committed by the Software for Pathology Messages in the Second Evaluation

Under-scrubbing Errors	Total	Missed
HIPAA-specified Identifiers		
Pathology reports (n = 7,193)	80,418	425 (0.5%)
Non–HIPAA-specified Identifiers		
Pathology reports (n = 7,193)	13,091	402 (3.1%)

HIPAA = Health Insurance Portability and Accountability Act.

Table 7 ■ Examples of Over-scrubbing Errors Committed by the Name Nearness Scrubber

Patient Name	Word Scrubbed
Brandon	random
Angel	range
Fred	red
Reed	red
Lora	flora
Ross	gross

set consisted of 220 discharge summaries. We did not introduce ambiguities in our test set. We processed real-world HL7 messages direct from our health information exchange. In the i2b2 challenge, ambiguities were intentionally introduced into the test set by replacing patient and provider names with medical terms. Finally, in our study MeDS had access to and used patient information in the header section of documents. The systems in the i2b2 study were not provided with equivalent information.

Overall, we found more non-HIPAA identifiers than HIPAA-specified identifiers, especially in laboratory reports. Approximately 70% of the non-HIPAA identifiers are provider names; the rest are laboratory/hospital names, addresses, and phone numbers. We describe these as non-HIPAA because provider information is not listed among the 19 variables (Figure 1) that HIPAA specifies as needing removal and is not strictly a patient identifier. However, most de-identification trials have removed provider information because it does provide information that could contribute to re-identification of the patient.

We performed several modifications to MeDS after our initial testing and achieved a modest improvement in HIPAA-specified identifier scrubbing performance. We added missing elements to several of the regular expressions (such as adding "suite" and "room #" to our regular expression that detects addresses), and we created several more regular expressions to detect more variations of date and accession number patterns. However, we found that the most significant modification needed to prevent missed identifiers was to adjust the order of processing by the regular expressions. Our initial evaluation showed that most missed patient identifier fragments were explained by a previous regular expression removing identifiers that caused subsequent regular expressions to be less effective. The following example illustrates this clearly. There is a regular expression that detects and scrubs any number over 4 digits. There is also a regular expression that detects a street address by looking for a pattern of "any number + any words + street identifier (i.e., street, road, boulevard, etc.)." If the number regular expression precedes the street address regular expression, errors can occur (i.e., "12345 Main Street" is converted to "xxx Main Street" prior to processing by the street address regular expression; therefore the "any number + any words + street identifier" pattern no longer exists). To prevent such errors, we discovered that, in general, during processing regular expressions that detect more specific patterns should precede those that detect more general patterns. Despite these errors, we deemed it very unlikely that a patient's identity could have been determined by any of these retained patient identifier fragments.

One of the strengths of our software is that it does not rely on a single method or process to remove identifiers. For example, a patient name in a report could be detected and scrubbed by the regular expression processor (i.e., the pattern "patient name: *firstname lastname*"), a direct match in the proper name list table, a match to the header information extracted earlier from the message, and finally by the word nearness similarity algorithm in the case of misspelled names. Perhaps any of the above processes alone would detect the patient name and the need to have multiple processes to remove a single kind of identifier may seem unnecessary. However, we found that this redundancy in the scrubbing processes lessens the likelihood of an identifier escaping detection.

Using data present in the header section of a message had a small effect on the overall accuracy of the scrubbing. Only a very small percentage of scrubbings depended on this technique. However, although this process is rarely needed for most reports, there could be instances when this process is invaluable, such as when a patient name is also a common word or a medical term. In such cases, the name is more likely to be missed by the pattern matching and name matching processes.

The ultimate goal of de-identification software is to scrub true patient identifiers while minimizing over-scrubbing. A medical report completely scrubbed of not only all patient identifiers but all important medical data as well is of no use to researchers. We considered that because we scrubbed information in excess of what HIPAA specifies and that our software committed 4,012 over-scrubbing errors, perhaps our scrubbed messages would no longer hold any research value. Therefore, we analyzed a sample of 300 scrubbed messages to determine readability and interpretability based on the following criteria. A laboratory message was interpretable if the type of test and the result was retained. A pathology report was interpretable if the type of report, specimen, and conclusion could be determined. A narrative report was interpretable if the majority of significant clinical data was retained, and the type of report and conclusion (if applicable) could be determined. An example of a scrubbed HL7 message is shown in Figure 5. Approximately 95% of scrubbed messages were both readable and interpretable.

The MeDS's name nearness scrubber committed many false-positive errors. Examples of these are shown in Table 7. Although clearly this process is not highly specific for detecting misspelled patient names, the importance of removing misspelled patient names in a report makes such a process valuable.

We acknowledge a limitation of de-identifying reports by removing HIPAA-specified patient identifiers. Despite removing the majority of the HIPAA-specified patient identifiers, data could occasionally remain that could potentially result in re-identification. Some documents, such as admission notes, typically contain detailed patient historical information. If a history is very unique, the identity of the patient could be compromised, especially when coupled with other data. The identities of "a former president of the United States with Alzheimer's disease" and "an HIV-positive, 6'9 inch black male, former professional basketball player" are probably readily apparent, despite the absence of any HIPAA identifiers. We did not find such occurrences in our dataset. This phenomenon illustrates the fact that although frequently the absence of patient identifiers is an adequate measure of de-identification, occasionally it is not. Eliminating from the dataset patient records of well-known individuals could help protect against such occurrences. In future versions of the software, we plan to add algorithms to scrub such contextual inferences.

There are several limitations to our study. The developer of the software also acted as the gold standard and evaluator of

```
MSH||CHARTLINC|SF_SOFTMED|||$YearMonthDay||||||||||||________75
PID|||$patient id$||$patient name$||||||||||||||||||
PV1|||||||$attending doctor$||$consulting doctor$||||||||
OBR|||||$requested date_time$|||||||||
OBX|
OBX|1|TX|<xxx> ^BEGIN DOCUMENT ^<xxx> ^<xxx> ^^<xxx>|| ENDOSCOPY REPORT DATE OF
PROCEDURE:<xxx>/<xxx> PROCEDURE:Upper <xxx> panendoscopy with Botox injections of the
pylorus. INDICATION:This is a <xxx>-year-old white male with diabetic gastroparesis who comes now
for Botox injection of the pylorus. ANESTHESIA:Versed 5 mg,Demerol 75 mg slowly
intravenously.Topical Topex was applied to the pharyngeal area. INSTRUMENT:Olympus <xxx>-130
gastroscope. PROCEDURE:With the patient in the left lateral decubitus position and after the above-
mentioned anesthesia was administered,the endoscope was passed under direct vision into the
esophagus.The esophagus appeared normal.The cardiac, fundus,body,antrum of the stomach including
retroflex views of the cardioesophageal junction and angularis incisura appeared normal.I did not see any
evidence of a bezoar of gastric distention.The pylorus,duodenal bulb,and second portion of the duodenum
were remarkable.The scope was withdrawn into the body of the stomach.There were a few antral
contractions. Then 200 units of Botox was injected in 1.5 to 2 mL increments in concentric circles around
the pylorus.He tolerated this without apparent complication. The scope was withdrawn and he tolerated the
procedure well. IMPRESSION:Successful Botox injection of the pylorus in a patient with gastroparesis.
DICTATED <xxx> <xxx>] <xxx> DICT:<xxx>/<xxx> <xxx> P #<xxx> <xxx> <xxx> <xxx>] <xxx>
DR.<xxx> <xxx> |||||| <xxx>||
```


Figure 5. Example of a scrubbed HL7 narrative report message (endoscopy report).

the scrubbing process. Ideally, several trained experts not part of the development team would perform the evaluation of the scrubbing process, allowing for interrator reliability measurements of agreement. All collected reports were part of the INPC network, which is limited to the central Indiana area. Processing reports originating from a different network in a different geographic area may affect scrubbing accuracy and may require software modification to achieve similar results.

Although these initial results are promising, we see several ways to improve our software. The addition of a geographic name database would lessen the possibility that portions of a patients address are missed, and has been used successfully in other de-identification systems.[9] We anticipate extending the name nearness scrubber to include other patient identifiers such as patient addresses, and provider names. Further modification to the name nearness scrubber is needed to lessen the likelihood of the software interpreting valid words as spelling errors of patient names. However, sophisticated natural language processing techniques would likely be needed for accurate determination of true spelling errors.

Conclusion

Our software successfully de-identified a wide range of medical documents from numerous sources and creates scrubbed reports that retain their interpretability, thereby maintaining their usefulness for clinical research. Occasional portions of HIPAA-specified patient identifiers missed by our software did not result in high risk of re-identification.

References ■

1. NOVA: Public Broadcasting System [homepage on the internet]. Louis Lasagna. Hippocratic Oath—Modern Version; 1964. Available from: http://www.pbs.org/wgbh/nova/doctors/oath_modern.html. Accessed July 17, 2008.
2. Tilton SH. Right to privacy and confidentiality of medical records. Occup Med 1996;11:17–29.
3. Kurtz G. EMR confidentiality and information security. J Healthc Inf Manag 2003;17:41–8.
4. Health and Human Services HIPAA Web site. Available at: http://www.hhs.gov/ocr/hipaa/. Accessed July 1, 2006.
5. U.S. Department of Health and Human Services. Standards for Privacy of Individually Identifiable Health Information; Final Rule. Code of Federal Regulations, Title 45, Parts 160 and 164. Available at: http://hhs.gov/ocr/combinedregtext.pdf. Accessed May 1, 2006.
6. Thomas SM, Mamlin B, Schadow G, McDonald C. A successful technique for removing names in pathology reports using an augmented search and replace method. Proc AMIA Symp 2002:777–81.
7. Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol 2004;121:176–86.
8. Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. Arch Pathol Lab Med 2003;127:680–6.
9. Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak 2006;6:12.
10. Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp 1996:333–7.
11. Sweeney L. Guaranteeing anonymity when sharing medical data, the Datafly System. Proc AMIA Annu Fall Symp 1997: 51–5.
12. Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. Medical document anonymization with a semantic lexicon. Proc AMIA Symp 2000:729–33.
13. De-id Web site. Available at: http://www.de-idata.com/. Accessed July 5, 2006.
14. Deidentify Web site. Available at: http://www.deidentify.com/. Accessed July 5, 2006.
15. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14: 550–63.
16. Wellner B, Huyck M, Mardis S, et al. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assoc 2007;14:564–73.
17. Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc 2007;14:574–80.
18. Health Level 7. HL7 Web site. Available at: http://www.hl7.org. Accessed July 1, 2006.
19. Henderson M. HL7 Messaging. 2nd ed. Aubrey, TX: O Tech-Health Care Technology Solutions; 2007.
20. McDonald CJ, Overhage JM, Tierney WM, et al. The Regenstrief Medical Record System: a quarter century experience. Int J Med Inform 1999;54:225–53.
21. McDonald CJ, Overhage JM, Barnes M, et al. The Indiana network for patient care: a working local health information

infrastructure. An example of a working infrastructure collaboration that links data from five health systems and hundreds of millions of entries. Health Aff (Millwood) 2005;24:1214–20.

22. Ispell Web site. Available at: http://ficus-www.cs.ucla.edu/geoff/ispell.html. Accessed May 1, 2006.
23. MetaMap Web site. Available at: http://mmtx.nlm.nih.gov/. Accessed July 5, 2006.
24. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1:161–74.
25. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11:392–402.
26. Friedman C, Sideli R. Tolerating spelling errors during patient validation. Comput Biomed Res 1992;25:486–509.